T-61.183 Special Course in Information Science III:
Support Vector Machines and Kernel Methods

Support Vector Regression

Based on Chapter 9 in *Learning with Kernels*
by B. Schölkopf and A.J. Smola

Pasi Lehtimäki
Pasi.Lehtimaki@hut.fi

Outline

- Introduction
- ε-SVR
- ν-SVR
- Applications
- Exercises

Introduction

- Classification problem: given training data $(x_1, y_1), \ldots, (x_m, y_m), x_i \in \mathcal{X}, y_i \in \{-1, 1\}$, estimate a function $f : \mathcal{X} \rightarrow \{-1, 1\}$, such that $y = f(x)$.
- Regression problem: given training data $(x_1, y_1), \ldots, (x_m, y_m), x_i \in \mathcal{X}, y_i \in \mathbb{R}$, estimate a real-valued function $f : \mathcal{X} \rightarrow \mathbb{R}$, such that $y = f(x)$.

Linear ε-SVR

- In the case of linear regression, the functions take the form

$$f(x) = \langle w, x \rangle + b$$

where $w, x \in \mathcal{H}$ and $b \in \mathbb{R}$.

- In this formulation, the data samples x and the parameters w are "already" located in the feature space $\mathcal{H}$ instead of mapping them from input space to the feature space using kernel functions.

- The optimization problem is to minimize

$$\frac{1}{2}\|w\|^2 + \frac{C}{m}\sum_{i=1}^{m} |y_i - f(x_i)|_\varepsilon,$$

where $|y_i - f(x_i)|_\varepsilon = \max\{0, |y - f(x)| - \varepsilon\}$is the ε-insensitive loss function.

- The previous optimization problem is equivalent to the minimisation of

$$\frac{1}{2}\|w\|^2 + \frac{C}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*)$$

subject to

$$(\langle w, x_i \rangle + b) - y_i \leq \varepsilon + \xi_i,$$

$$y_i - (\langle w, x_i \rangle + b) \leq \varepsilon + \xi_i^*,$$

$$\xi_i^{(*)} \geq 0.$$

- In this formulation, the prediction errors greater than ε are included in *slack* variables ξ_i (predicted output greater than true output) and ξ_i^* (predicted output less than true output).

- The training error gets minimised in term $\frac{C}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*)$ and the complexity gets minimised in term $\frac{1}{2}\|w\|^2$.
- Why to minimise the norm of the weight vector w?
- In classification, the minimisation of the $\|w\|$ corresponds to maximisation of the margin of separation $\rho = 2/\|w\|$.
- In regression, it corresponds to maximisation of the so-called ε-margin $m_\varepsilon(f)$ of a function f:

$$m_\varepsilon(f) := \inf\{\|x - x'\| \mid x, x \in \mathcal{X}, \|f(x) - f(x')\| \geq 2\varepsilon\}.$$

- Thus, the *flatness* of the decision surface is maximised. In other words, the distance of the closest two points in the feature space for which the predicted output differs at least 2ε is maximized. As a result, “small variation in x produces small variation in y”.

The dual formulation for linear ε-SVR

- The partial derivatives of the Lagrangian of the primal objective function w.r.t primal variables $(w, b, \xi_i^{(*)})$ must be equal to zero, leading to the dual optimization problem

$$\begin{aligned}
\text{maximize} \qquad & -\frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i-\alpha_i^*)(\alpha_i-\alpha_i^*)\left\langle x_i, x_j\right\rangle \\
& -\varepsilon\sum_{i=1}^{m}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{m}y_i(\alpha_i-\alpha_i^*) \\
\text{subject to} \qquad & \sum_{i=1}^{m}(\alpha_i-\alpha_i^*)=0, \\
& \alpha_i, \alpha_i^* \in [0, C/m]
\end{aligned}$$

- The estimated function f becomes

$$f(x) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b,$$

where the primal variables w and b are computed using

$$\begin{aligned} w &= \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) x_i \\ b &= y_i - \langle w, x_i \rangle - \varepsilon \quad \text{for} \quad \alpha_i \in (0, C/m) \\ b &= y_i - \langle w, x_i \rangle + \varepsilon \quad \text{for} \quad \alpha_i^* \in (0, C/m) \end{aligned}$$

- Only points (x_i, y_i) with $\alpha_i^{(*)} = C/m$ are outside ε-tube.
- For $\alpha_i^{(*)} \in (0, C/m)$, we have $\xi_i^{(*)} = 0$.
- For all i, $\alpha_i \alpha_i^* = 0$.
- Points with nonvanishing α_i, α_i^* are called *support vectors*.

Nonlinear ε-SVR

- Nonlinear function f is obtained by mapping training samples x from input space $\mathcal{X}$ to the feature space $\mathcal{H}$ via a nonlinear map $\Phi(x)$ and solving a linear regression problem in the feature space.
- In such case, the inner product appearing in the dual optimization problem becomes

$$\langle \Phi(x), \Phi(x') \rangle = K(x, x')$$

- Again, the nonlinear map $\Phi(x)$ is not explicitly needed in optimization.

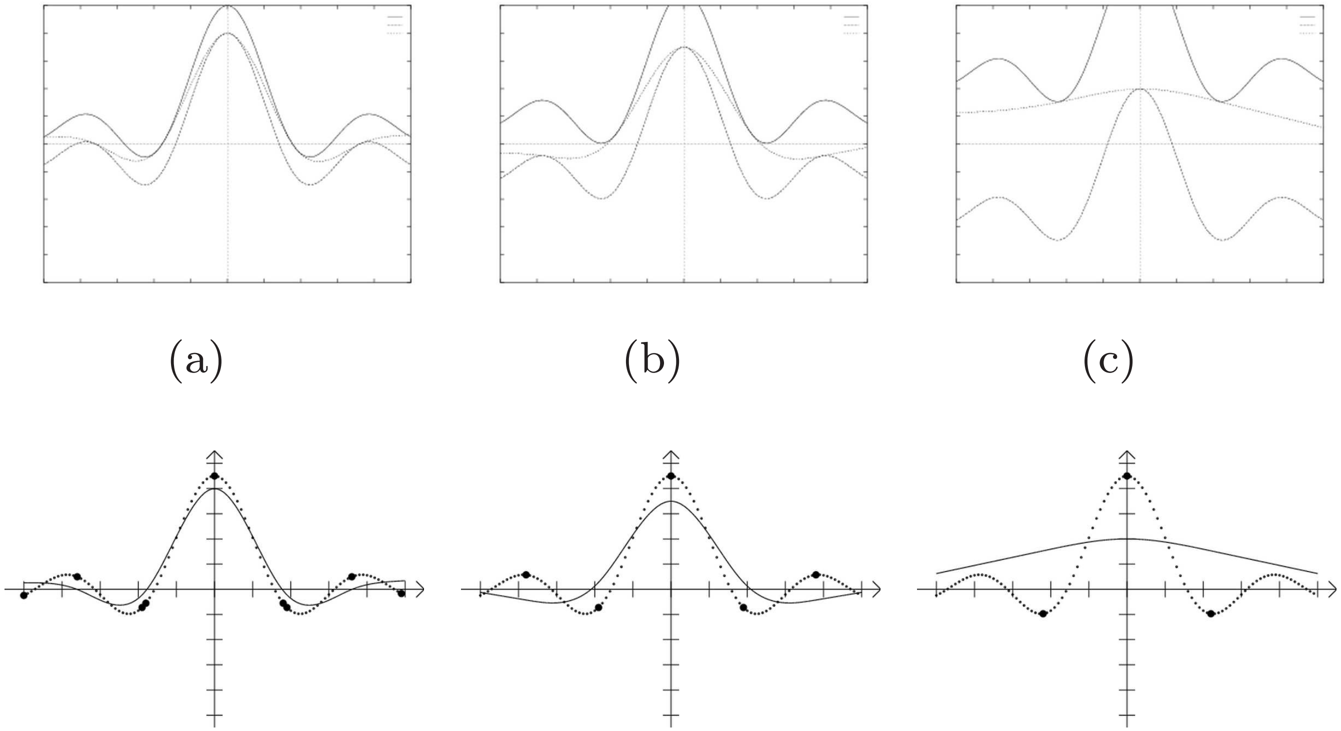
- The dual optimization problem becomes

$$
\begin{aligned}
\text{maximize} \qquad & -\frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i - \alpha_i^*)(\alpha_i - \alpha_i^*)K(x_i, x_j) \\
& -\varepsilon \sum_{i=1}^{m}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i - \alpha_i^*) \\
\text{subject to} \qquad & \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0, \\
& \alpha_i, \alpha_i^* \in [0, C/m],
\end{aligned}
$$

- The estimated nonlinear function becomes

$$
f(x) = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)K(x_i, x) + b
$$

Example:



(a) (b) (c)

(d) (e) (f)

Figure 1: In (a,d) $\varepsilon = 0.1$, in (b,e) $\varepsilon = 0.2$, and in (c,f) $\varepsilon = 0.5$

Linear ν-SVR

- In ε-SVR, the user had to select both the value of ε and C.
- In ν-SVR, the accuracy of approximation ε is computed automatically to be as accurate as possible.
- The objective function

$$\frac{1}{2}\|w\|^2 + C\left(\nu\varepsilon + \frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*)\right),$$

where $\nu \geq 0$, is minimized subject to

$$(\langle w, x_i\rangle + b) - y_i \leq \varepsilon + \xi_i,$$

$$y_i - (\langle w, x_i\rangle + b) \leq \varepsilon + \xi_i^*,$$

$$\xi_i^{(*)} \geq 0,$$

$$\varepsilon \geq 0.$$

The dual of linear ν-SVR

- The dual is obtained similarly to the previous cases: form the Lagrangian of the objective function and set the partial derivatives w.r.t the primal variables $(w, b, \varepsilon, \xi_i^{(*)})$ equal to zero. This leads to

$$
\begin{aligned}
\text{maximize} \quad & -\frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i-\alpha_i^*)(\alpha_i-\alpha_i^*)\left\langle x_i, x_j\right\rangle+\sum_{i=1}^{m} y_i(\alpha_i-\alpha_i^*) \\
\text{subject to} \quad & \sum_{i=1}^{m}(\alpha_i-\alpha_i^*)=0, \\
& \alpha_i, \alpha_i^* \in [0, C/m], \\
& \sum_{i=1}^{m}(\alpha_i+\alpha_i^*) \leq C\nu
\end{aligned}
$$

Nonlinear ν-SVR

- The nonlinear ν-SVR can be obtained similarly to the ε-SVR case by replacing the dot product $\langle x_i, x_j \rangle$ with the kernel $K(x_i, x_j)$ selected by the user.
- The resulting nonlinear regression function is similar to the ε-SVR case:

$$f(x) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) K(x_i, x) + b,$$

  where b is computed similarly as in the ε-SVR.
- Also, the optimal value for ε can be computed from the same conditions as b.

The selection of ν

- The only sensible values for ν are $0 \geq \nu \geq 1$ (explained later).
- ν is an upper bound on the fraction of errors (examples outside the ε-tube):
  - imagine ε that grows from zero to a large number.
  - the loss function decreases as long as the latter term (average loss) is larger than ν, that is, when $1/m \sum_i (\xi_i - \xi_i^*) \geq \nu$.
  - Thus, the optimum must be when $1/m \sum_i (\xi_i - \xi_i^*) \leq \nu$.
  - Thus, ν is an upper bound for the average amount of slack.
  - Since slack ξ_i^* is only produced by the training samples outside the ε-tube, ν is an upper bound for the fraction of errors.

- ν is a lower bound for the fraction of support vectors (examples outside the ε-tube and on the edge of the ε-tube):
  - imagine large ε that shrinks towards zero.
  - we want small loss: the first term in loss function decreases and the second one increases.
  - this continues as long as $\nu \geq 1/m \sum_i (\xi_i - \xi_i^*)$.
  - the optimum must therefore be when $\nu \leq 1/m \sum_i (\xi_i - \xi_i^*)$.
  - This time, ν is a lower bound for the average amount of slack.
  - Therefore, there must be at least certain number of training samples to produce the minimum amount of slack, that is, enough training samples outside the ε-tube.
  - Since ν is a lower bound for the fraction of points outside the ε-tube, it must be a lower bound for the fraction of support vectors also.

- For both fractions of the bounds, only sensible values are $0 \geq \nu \geq 1$.
- If ν-SVR has the solution $\bar{\varepsilon}, \bar{w}, \bar{b}$, then ε-SVR with $\varepsilon = \bar{\varepsilon}$ has the solution $\bar{w}, \bar{b}$.
- In SVR with ε-insensitive loss function, local movements of y_i outside the ε-tube do not influence the regression.

Example:

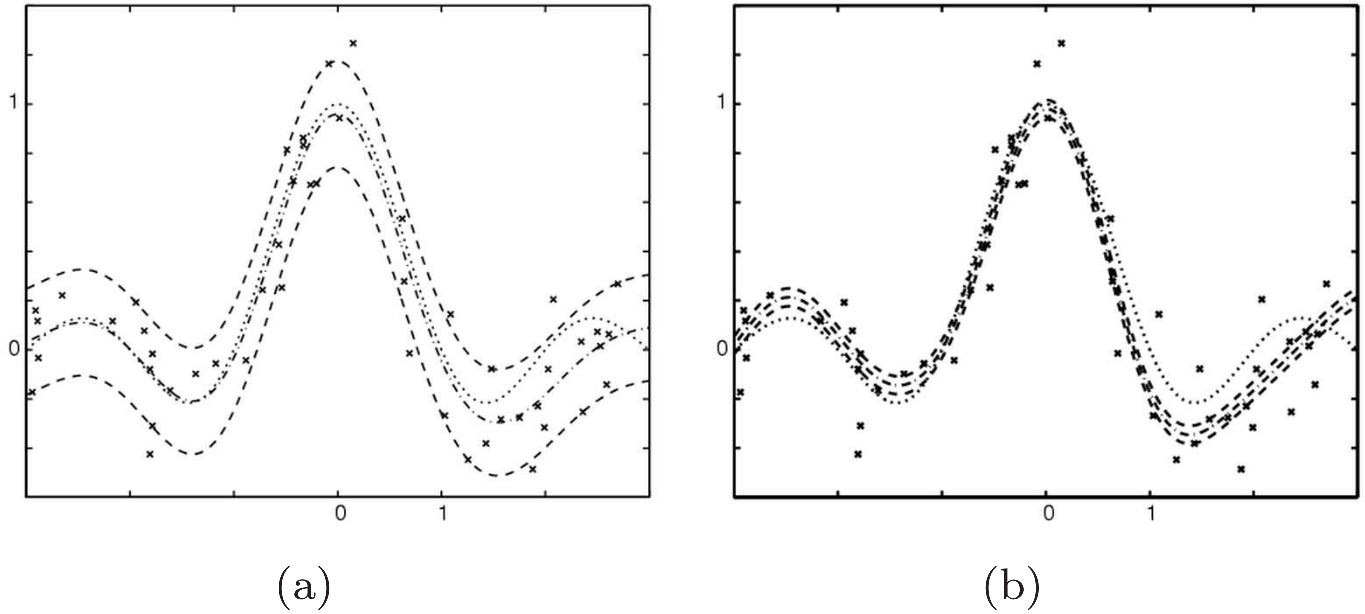


Figure 2: (a) $\nu = 0.2$ results with $\varepsilon = 0.22$ and (b) $\nu = 0.8$ results with $\varepsilon = 0.04$

Application: Boston Housing

ν	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
automatic ε	2.6	1.7	1.2	0.8	0.6	0.3	0.0	0.0	0.0
MSE	9.4	8.7	9.3	9.5	10.0	10.6	11.3	11.3	11.3
Errors	0.0	0.1	0.2	0.2	0.3	0.4	0.5	0.5	0.5
SVs	0.3	0.4	0.6	0.7	0.8	0.9	1.0	1.0	1.0

ε	0	1	2	3	4	5	6	7	8	9
MSE	11.3	9.5	8.8	9.7	11.2	13.1	15.6	18.2	22.1	27.0
Errors	0.5	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVs	1.0	0.6	0.4	0.3	0.2	0.1	0.1	0.1	0.1	0.1

- For ν-SVR, when $\nu \geq 0.7$, the ε is tuned to zero. The results are the same as for fixing the ε to zero in ε-SVR.

Conclusion

- Two SVM methods for solving regression problems were introduced.
- In these methods, the only difference was in the user-selected parameters, which had quite a different interpretation.
- The SVM for regression is an efficient tool when trying to achieve good generalisation on new samples, since it minimizes both the training error and the complexity of the model simultaneously, and the solution for nonlinear regression can be obtained with quadratic programming techniques, providing a global solution.
- Exercises: Problem 9.1 and Problem 9.2.